Mathematical Optimization in Radiotherapy Treatment Planning

Ehsan Salari

Department of Radiation Oncology
Massachusetts General Hospital and Harvard Medical School

HST S14
May 6, 2013

Outline

1. Intensity-modulated Radiotherapy (IMRT)
2. Sequential Method: Fluence-map Optimization (FMO)
3. Sequential Method: Leaf Sequencing (LS)
4. Direct Aperture Optimization (DAO)

Intensity-modulated Radiotherapy (IMRT)

- In 3D-CRT, radiation fluence across the opening area of the aperture is constant
- To better spare organs-at-risk more *fluence modulation* is needed
- *Intensity-modulated radiotherapy* (IMRT) is a more recent modality that allows for more fluence modulation at each beam

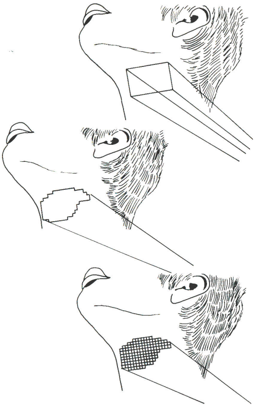
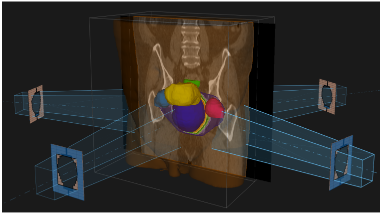
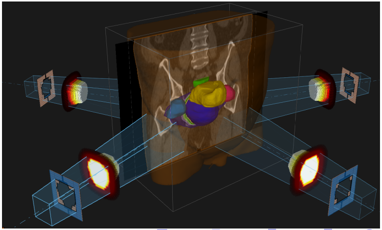


Figure: [Webb, 2001]

Comparing 3D-CRT and IMRT

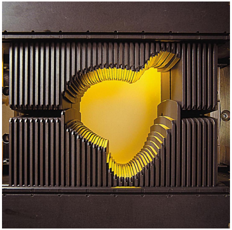
- 3D-CRT shapes apertures that conform to tumor shape
- IMRT creates a *fluence map* (intensity profile) per beam

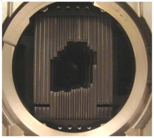




Multi-leaf Collimator (MLC)

- In IMRT
  - gantry head is equipped with a *multi-leaf collimator* (MLC) system
  - MLC leaves form apertures with different shapes and intensities





Creating Fluence Maps using MLC

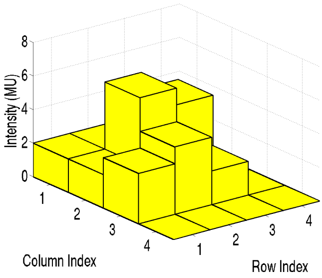
- Using MLC a desired fluence map can be created

2	2	0	0
2	2	2	0
0	0	0	0
0	0	0	0

0	0	0	0
0	1	1	0
0	1	0	0
0	0	0	0

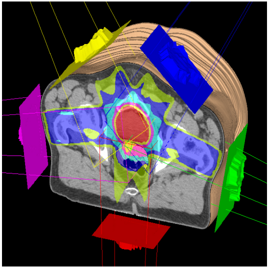
0	0	0	0
0	3	0	0
3	3	0	0
0	0	0	0

0	0	0	0
0	0	2	0
0	0	2	0
0	0	0	0



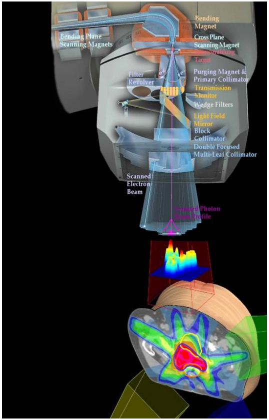
IMRT Treatment Planning

- IMRT planning is to determine a set of apertures and their intensities that yield a dose distribution that
  - adequately covers *target(s)*
  - preserves functionality of *critical structures*



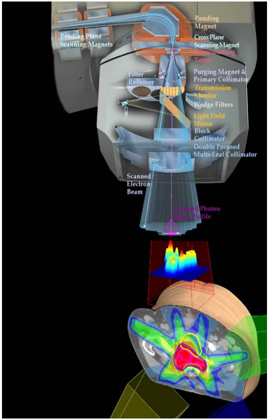
Solution Approaches to IMRT Treatment Planning

- *Sequential method*
  - (1) *Beam orientation optimization* (BOO)
    - determines a set of beam directions
    - is usually performed manually
  - (2) *Fluence-map optimization* (FMO)
    - determines an intensity profile for each beam
  - (3) *Leaf sequencing* (LS)
    - decomposes intensity profiles to deliverable apertures
- *Direct aperture optimization* (DAO)
  - integrates FMO and LS



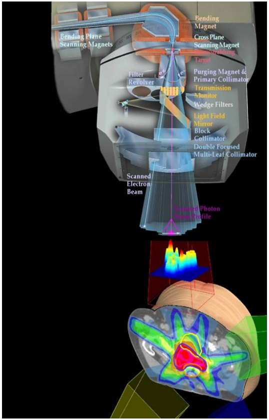
Solution Approaches to IMRT Treatment Planning

- *Sequential method*
  - **(1)** *Beam orientation optimization* (BOO)
    - determines a set of beam directions
    - is usually performed manually
  - **(2)** *Fluence-map optimization* (FMO)
    - determines an intensity profile for each beam
  - **(3)** *Leaf sequencing* (LS)
    - decomposes intensity profiles to deliverable apertures
- *Direct aperture optimization* (DAO)
  - integrates FMO and LS



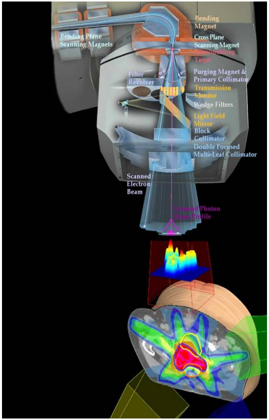
Solution Approaches to IMRT Treatment Planning

- *Sequential method*
  - **(1)** *Beam orientation optimization* (BOO)
    - determines a set of beam directions
    - is usually performed manually
  - **(2)** *Fluence-map optimization* (FMO)
    - determines an intensity profile for each beam
  - **(3)** *Leaf sequencing* (LS)
    - decomposes intensity profiles to deliverable apertures
- *Direct aperture optimization* (DAO)
  - integrates FMO and LS



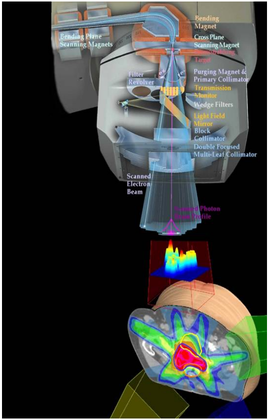
Solution Approaches to IMRT Treatment Planning

- *Sequential method*
  - **(1)** *Beam orientation optimization* (BOO)
    - determines a set of beam directions
    - is usually performed manually
  - **(2)** *Fluence-map optimization* (FMO)
    - determines an intensity profile for each beam
  - **(3)** *Leaf sequencing* (LS)
    - decomposes intensity profiles to deliverable apertures
- *Direct aperture optimization* (DAO)
  - integrates FMO and LS



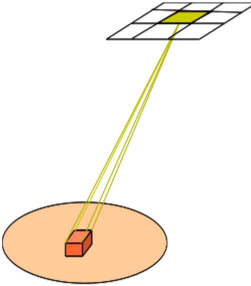
Solution Approaches to IMRT Treatment Planning

- *Sequential method*
  - **(1)** *Beam orientation optimization* (BOO)
    - determines a set of beam directions
    - is usually performed manually
  - **(2)** *Fluence-map optimization* (FMO)
    - determines an intensity profile for each beam
  - **(3)** *Leaf sequencing* (LS)
    - decomposes intensity profiles to deliverable apertures
- *Direct aperture optimization* (DAO)
  - integrates FMO and LS



Fluence-map Optimization (FMO)

- Rectangular beams are discretized into *beamlets* $i \in I$
- Using pencil-beam dose calculation method, *beamlet dose deposition coefficients* $D = [D_{iv}]$ $(i \in I, v \in V)$ are computed
- Using optimization methods, optimal fluence maps x_i $(i \in I)$ are determined
  - large-scale problem: $\mathcal{O}\left(10^3\right)$ beamlets and $\mathcal{O}\left(10^5\right)$ voxels



FMO Mathematical Formulation

- Mathematical formulation for the FMO problem

$$\min G(\mathbf{d})$$

subject to

$$\begin{aligned} \mathbf{d} &= D^\top \mathbf{x} \\ H(\mathbf{d}) &\leq 0 \\ \mathbf{x} &\geq \mathbf{0} \end{aligned}$$

- Notation
  - $\mathbf{x}$: vector of beamlet intensities
  - $D = [D_{iv}]$: matrix of beamlet dose deposition coefficients

NLP Solution Approach for FMO

- FMO problem can be solved using *interior-point method* (*barrier method*)
  - see [Bazaraa et al., 2006]
  - transform the constrained problem to unconstrained problem using barrier function
  - sets a barrier against leaving the feasible region

$$\min_{\mathbf{x}} \; G\left(D^\top \mathbf{x}\right) + \mu \underbrace{B(\mathbf{x})}_{\text{barrier}} \qquad \mu > 0$$

- Barrier function characteristics are
  - nonnegative and continuous over $\{\mathbf{x} : \mathbf{x} \geq \mathbf{0}, H\left(D^\top \mathbf{x}\right) \leq \mathbf{0}\}$
  - approaches ∞ as $\mathbf{x}$ approaches the boundary from interior

Barrier Method for FMO

- We formulate a parametric problem

$$\phi(\mu) = \min_{\mathbf{x}} \; G\left(D^\top \mathbf{x}\right) + \mu B(\mathbf{x})$$

- It can be shown that

$$\lim_{\mu \to 0^+} \phi(\mu) = \min_{\mathbf{x}} \left\{ G\left(D^\top \mathbf{x}\right) : \mathbf{x} \geq \mathbf{0}, H\left(D^\top \mathbf{x}\right) \leq \mathbf{0} \right\}$$

Barrier Method for FMO

1. Initialize: choose interior point $\mathbf{x}_0 > \mathbf{0}$, $\mu_0 > 0$, and $0 < \beta < 1$
2. Main step: at iteration k solve unconstrained problem

$$\min_{\mathbf{x}} \; G\left(D^\top \mathbf{x}\right) + \mu_k B(\mathbf{x})$$

   to obtain optimal solution $\mathbf{x}_k$
3. Termination condition: if $\mu_k B(\mathbf{x}_k) < \epsilon$, stop; otherwise, $\mu_{k+1} = \beta \mu_k$ and go to step 2

Unconstrained Optimization

- Consider unconstrained optimization problem

$$\min \; F(\mathbf{x})$$

- $\bar{\mathbf{x}}$ is a *global minimum* if $F(\bar{\mathbf{x}}) \leq F(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$
- $\bar{\mathbf{x}}$ is a *local minimum* if there is an ϵ-*neighborhood* $N_\epsilon(\bar{\mathbf{x}})$ around $\bar{\mathbf{x}}$ such that $F(\bar{\mathbf{x}}) \leq F(\mathbf{x})$ for all $\mathbf{x} \in N_\epsilon(\bar{\mathbf{x}})$
  - we assume differentiability
  - see [Bazaraa et al., 2006]

Characterizing Local Minimum

- $\mathbf{s}$ is a *descent* direction at $\bar{\mathbf{x}}$ if

$$\lim_{\lambda \to 0^+} \frac{F(\bar{\mathbf{x}} + \lambda \mathbf{s}) - F(\bar{\mathbf{x}})}{\lambda} = \nabla F(\bar{\mathbf{x}})^\top \mathbf{s} < 0$$

- Necessary condition: if $\bar{\mathbf{x}}$ is a local minimum, then $\nabla F(\bar{\mathbf{x}}) = \mathbf{0}$
- Sufficient condition: if $\nabla F(\bar{\mathbf{x}}) = \mathbf{0}$ and $\nabla^2 F(\bar{\mathbf{x}}) \succ 0$, then $\bar{\mathbf{x}}$ is a local minimum

Class of Convex Functions

- *Convex* functions
  - Definition $\forall\, \bar{\mathbf{x}}, \hat{\mathbf{x}} \in \mathbb{R}^n$

$$F\left(\lambda \bar{\mathbf{x}} + (1-\lambda)\hat{\mathbf{x}}\right) \leq \lambda F\left(\bar{\mathbf{x}}\right) + (1-\lambda) F\left(\hat{\mathbf{x}}\right) \qquad \lambda \in (0,1)$$

  - F is convex if and only if $\nabla^2 F$ is positive semi-definite everywhere
- If F is convex, then $\bar{\mathbf{x}}$ is a global minimum if and only if $\nabla F\left(\bar{\mathbf{x}}\right) = \mathbf{0}$
  - a desired property for unconstrained optimization

Steepest Descent for Unconstrained Optimization

- Starting from $\bar{\mathbf{x}}$ it iteratively moves toward local minimum
- *Steepest descent* at $\bar{\mathbf{x}}$ can be obtained by

$$\min \nabla F(\bar{\mathbf{x}})^\top \mathbf{s}$$

subject to

$$\|\mathbf{s}\| \leq 1$$

- which yields

$$\mathbf{s} = -\frac{\nabla F(\bar{\mathbf{x}})}{\|\nabla F(\bar{\mathbf{x}})\|}$$

Steepest Descent for Unconstrained Optimization

1. Initialize: Let $\epsilon > 0$, choose starting point $\mathbf{x}_0$
2. Steepest descent direction: At iteration k, let

$$\mathbf{s}_k = -\frac{\nabla F(\mathbf{x}_k)}{\|\nabla F(\mathbf{x}_k)\|}$$

3. Termination condition: If $\|\mathbf{s}_k\| < \epsilon$ stop; else, go to step 4
4. Line search:

$$\lambda^* = \operatorname*{argmin}_{\lambda \geq 0} F(\mathbf{x}_k + \lambda \mathbf{s}_k)$$

5. Update solution: $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda^* \mathbf{s}_k$ and go to step 2

Line Search

- *Line search* is to find optimal step length to move from point **x** along direction **s**
- It is rarely possible to obtain analytical solutions

$$\frac{\partial F(\mathbf{x}+\lambda\mathbf{s})}{\partial \lambda} = \mathbf{s}^\top \nabla F(\mathbf{x}+\lambda\mathbf{s}) = 0$$

- Numerical methods are commonly used
  - see [Bazaraa et al., 2006]

Line Search: Uncertainty Interval

- Derivative-free numerical solution method for

$$\min_{a \leq \lambda \leq b} \theta(\lambda) = F(\mathbf{x} + \lambda \mathbf{s})$$

- To reduce *uncertainty interval* $[a, b]$ we evaluate $\theta(\lambda)$ for different $\lambda \in [a, b]$
- Suppose θ is *strictly quasi-convex* (unimodal). Let $\lambda_1, \lambda_2 \in [a, b]$
  - If $\theta(\lambda_1) \leq \theta(\lambda_2)$, then $\theta(\lambda) \geq \theta(\lambda_1)$ for $\lambda \in [\lambda_2, b]$
  - If $\theta(\lambda_1) \geq \theta(\lambda_2)$, then $\theta(\lambda) \geq \theta(\lambda_2)$ for $\lambda \in [a, \lambda_1]$

Line Search: Dichotomous Search

1. Initialize: set initial uncertainty interval $[a_0, b_0]$, distinguishing param. $2\epsilon > 0$, and threshold param. δ
2. Main step: let $\lambda_1 = \frac{a_k+b_k}{2} - \epsilon$ and $\lambda_2 = \frac{a_k+b_k}{2} + \epsilon$, then

$$[a_{k+1}, b_{k+1}] = \begin{cases} [a_k, \lambda_2] & \text{if } \theta(\lambda_1) \leq \theta(\lambda_2) \\ [\lambda_1, b_k] & \text{otherwise} \end{cases}$$

3. Termination condition, if $b_{k+1} - a_{k+1} < \delta$ stop; otherwise, go to Step 2

Example of FMO Mathematical Formulation

- Dose evaluation criteria: summation of piecewise quadratic voxel-based penalties for all relevant structures $s \in S$

$$\min_{\mathbf{x}\geq\mathbf{0}} G\left(D^\top\mathbf{x}\right) = \sum_{s\in S}\sum_{v\in V_s} \gamma_s^+ \underbrace{\max\left\{\sum_{i\in I} D_{iv}x_i - t_v, 0\right\}^2}_{\text{overdosing penalty}}$$

$$+\gamma_s^- \underbrace{\max\left\{t_v - \sum_{i\in I} D_{iv}x_i, 0\right\}^2}_{\text{underdosing penalty}}$$

- We assume only nonnegativity constraints $\mathbf{x} \geq \mathbf{0}$, results can be generalized to include dose constraints

Example of FMO Mathematical Formulation

- *Logarithmic barrier* function for nonnegativity of beamlet intensities

$$\phi(\mu) = \min_{\mathbf{x}} \; G\left(D^\top \mathbf{x}\right) - \mu \sum_{i \in I} \ln(x_i)$$

- We solve $\phi(\mu)$ for $\mu > 0$ using Steepest Descent method
- Alternatively we can use *primal-dual interior-point method*
  - see [Aleman et al., 2010]

Primal-dual Interior Point Method

- To obtain $\phi(\mu)$ we find $\mathbf{x}^*$ where gradient vanishes

$$\frac{\partial\, G\left(D^\top \mathbf{x}\right) - \mu \sum_{i\in I} \ln\left(x_i\right)}{\partial\, x_i} = \frac{\partial\, G\left(D^\top \mathbf{x}\right)}{\partial\, x_i} - \frac{\mu}{x_i} = 0 \quad i \in I$$

- Variable transformation (x, λ: primal and dual variables)

$$\lambda_i = \frac{\mu}{x_i} \qquad i \in I$$

- Solve nonlinear system of equations for $\mathbf{x}, \boldsymbol{\lambda}$

$$\nabla_{\mathbf{x}} G - \boldsymbol{\lambda} = \mathbf{0}$$
$$\Lambda \cdot X = \mu\, \mathbf{e}$$

Primal-dual Interior Point Method

- Newton method to solve nonlinear system of equation
  1. Main step: determine direction and step length

$$\begin{pmatrix} \Delta \mathbf{x}_{(k)} \\ \Delta \boldsymbol{\lambda}_{(k)} \end{pmatrix} = -\begin{pmatrix} \nabla^2_{\mathbf{xx}} G_{(k)} & -I \\ \Lambda_{(k)} & X_{(k)} \end{pmatrix}^{-1} \begin{pmatrix} \nabla_{\mathbf{x}} G_{(k)} - \boldsymbol{\lambda} \\ \Lambda_{(k)} \cdot X_{(k)} - \mu_{(k)} \mathbf{e} \end{pmatrix}$$

  2. Update solution

$$\begin{pmatrix} \mathbf{x}_{(k+1)} \\ \boldsymbol{\lambda}_{(k+1)} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{(k)} \\ \boldsymbol{\lambda}_{(k)} \end{pmatrix} + \alpha_{(k)} \begin{pmatrix} \Delta \mathbf{x}_{(k)} \\ \Delta \boldsymbol{\lambda}(k) \end{pmatrix}$$

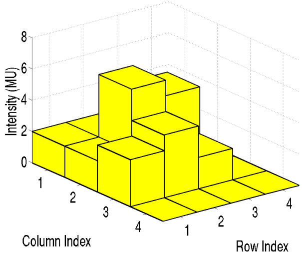
  3. Termination condition: if $\mathbf{x}^{\top}_{(k+1)} \boldsymbol{\lambda}_{(k+1)} < \epsilon$, then stop; otherwise go to Step 1

Leaf Sequencing Problem

- How to decompose fluence map into collection of deliverable apertures?
  - we assume *step-and-shoot* delivery
  - apertures are *binary matrices* with *consecutive ones* at each row

$$2\times\begin{bmatrix}1&1&0&0\\1&1&1&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\quad 1\times\begin{bmatrix}0&0&0&0\\0&1&1&0\\0&1&0&0\\0&0&0&0\end{bmatrix}$$

$$3\times\begin{bmatrix}0&0&0&0\\0&1&0&0\\1&1&0&0\\0&0&0&0\end{bmatrix}\quad 2\times\begin{bmatrix}0&0&0&0\\0&0&1&0\\0&0&1&0\\0&0&0&0\end{bmatrix}$$



Leaf Sequencing (LS)

- There is a large number of possible decompositions
- *Leaf Sequencing* (LS) aims at finding decomposition with
  - minimal total monitor units (*beam-on time*)
  - minimal number of binary matrices
  - total treatment time depends on beam-on time and number of apertures
- Assumptions
  - there is only *row-convexity* constraint on aperture shapes
    - see [Baatar et al., 2005] for additional MLC hardware constraints
  - integral intensities by rounding fluence map X

LS Formulation: Minimizing Beam-on Time

- Beam-on time minimization

$$\min \sum_{k \in K} \alpha_k$$

subject to

$$X = \sum_{k \in K} \alpha_k S_k$$

$$\alpha_k \geq 0 \qquad\qquad k \in K$$

- Notation
  - K: set of $M \times N$ binary matrices with consecutive-ones property at each row
  - S_k: binary matrix $k \in K$
  - α_k: number of MU for binary matrix $k \in K$

Minimizing Beam-on Time: Example

- Example

$$\sum_{k \in K} \alpha_k = 5$$

$$\underbrace{\begin{pmatrix} 2 & 5 & 3 \\ 3 & 4 & 2 \end{pmatrix}}_{X} = \underbrace{2\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}}_{\alpha_1 S_1} + \underbrace{1\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}}_{\alpha_2 S_2} + \underbrace{2\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}}_{\alpha_3 S_3}$$

  - S_1, S_2, S_3 have positive monitor units
  - $\alpha_k = 0$ for all other binary matrices $k \in K$

Minimizing Beam-on Time: Solution Approach

- Fluence map

$$X = \begin{pmatrix} 2 & 5 & 3 \\ 3 & 4 & 2 \end{pmatrix}$$

- Difference matrix

$$\tilde{X} = \left[\ X_{m,n} - X_{m,n-1}\ \right]_{M\times(N+1)} = \begin{pmatrix} 2 & 3 & -2 & -3 \\ 3 & 1 & -2 & -2 \end{pmatrix}$$

- Whenever $\tilde{X}_{m,n} > 0$, decomposition needs to use interval with left boundary in bixel m with at least $\tilde{X}_{m,n}$ MU.

Minimizing Beam-on Time: Solution Approach

- Sum of positive gradient (SPG)

$$SPG_m = \sum_{n=1}^{N+1} \max\left\{\tilde{X}_{m,n}, 0\right\} = \sum_{n=0}^{N} \max\left\{0, X_{m,n+1} - X_{m,n}\right\}$$

$$SPG(X) = \max_m \{SPG_m\} = \max\{4, 5\}$$

- $SPG(X)$ provides a lower bound on the required beam-on time for delivering X
  - there are decompositions for which LB can be obtained

Minimizing Beam-on Time: Solution Approach

- One can decompose each fluence row individually

$$X = \begin{pmatrix} 2 & 5 & 3 \\ 3 & 4 & 2 \end{pmatrix}$$

$$X_{1\cdot} = \begin{pmatrix} 2 & 5 & 3 \end{pmatrix} \rightarrow \sum_{k \in K} \alpha_{1k} V_k$$

$$X_{2\cdot} = \begin{pmatrix} 3 & 4 & 2 \end{pmatrix} \rightarrow \sum_{k \in K} \alpha_{2k} V_k$$

- Row decompositions can be combined to form LS with minimal beam-on time

$$\alpha_{k''} S_{k''} = \min \{\alpha_{1k}, \alpha_{2k'}\} \begin{pmatrix} V_k \\ V_{k'} \end{pmatrix}$$

Minimizing Beam-on Time: Network-flow Model

- Decomposing fluence row m: $(X_{mn} : n \in N)^{\top}$
  - see [Ahuja and Hamacher, 2005]
- Let K be the collection of all binary vectors with consecutive-ones property

$$\min \sum_{k \in K} \alpha_k$$

subject to

$$\begin{pmatrix} 1 & 0 & 0 & 1 & \dots & 1 \\ 0 & 1 & 0 & 1 & \dots & 1 \\ 0 & 0 & 1 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{pmatrix} = \begin{pmatrix} X_{m1} \\ X_{m2} \\ \vdots \\ X_{mN} \end{pmatrix}$$

$$\alpha_k \geq 0 \qquad \forall k \in K$$

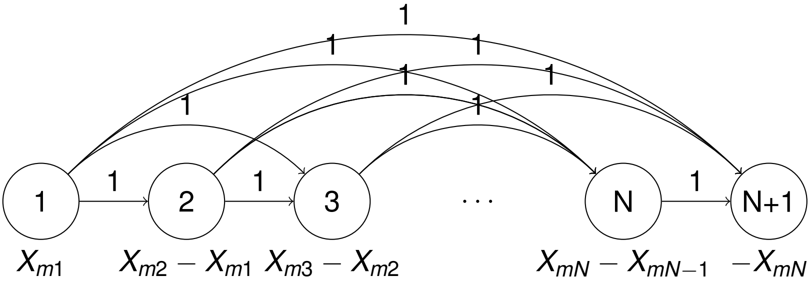
Minimizing Beam-on Time: Network-flow Model

- It can be represented using a *network-flow* model
  - by adding a zero vector at the end of constraint matrix and subtracting row n from n+1

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 1 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ 0 & 0 & -1 & \dots & -1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{pmatrix} = \begin{pmatrix} X_{m1} \\ X_{m2} - X_{m1} \\ \vdots \\ X_{mN} - X_{mN-1} \\ -X_{mN} \end{pmatrix}$$

Minimizing Beam-on Time: Network-flow Model

- Network representation
  - nodes are bixels $n = 1, \ldots, N + 1$
  - arcs are binary vectors with consecutive ones $k \in K$
  - nodes have supply/demand
- What is the minimum-cost flow to satisfy all node demands?



Network-flow Model: Min-cost Flow Algorithm

- Surplus/demand for node $n = 1, \ldots, N$ is defined as

$$b(n) = X_{mn} - X_{mn-1}$$

- Flow is sent from nodes with surplus $b(n) > 0$ to nodes with demand $b(n') < 0$

1. Initialize: $u_0 = \min\{n : b_n > 0\}$, $v_0 = \min\{n : b_n < 0\}$
2. Main step: at iteration k

$$\alpha_{(u_k,v_k)} = \min\{b(u_k), -b(v_k)\}$$
$$b(u_k) \leftarrow b(u_k) - \alpha_{(u_k,v_k)}$$
$$b(v_k) \leftarrow b(v_k) + \alpha_{(u_k,v_k)}$$

3. Termination condition: if $u_k, v_k = N + 1$, stop; else, increment accordingly

Minimizing Number of Apertures

- LS with objective of minimizing number of apertures is more involved
  - belongs to the class of NP-hard problems (see [Baatar et al., 2005])
    - in contrast with minimizing beam-on time which can be solved in polynomial time
  - fluence rows cannot be decomposed individually
- Solution approaches seek for decompositions using minimal number of apertures while constraining beam-on time to SPG

LS Formulation: Minimizing Number of Apertures

- Minimizing number of apertures

$$\min \|\boldsymbol{\alpha}\|_0$$

subject to

$$X = \sum_{k \in K} \alpha_k S_k$$

$$\sum_{k \in K} \alpha_k \leq SPG(X)$$

$$\alpha_k \geq 0 \qquad k \in K$$

- Notation
  - K: set of $M \times N$ binary matrices with consecutive-ones property at each row
  - S_k: binary matrix $k \in K$
  - α_k: number of MU for binary matrix $k \in K$

Minimizing Number of Apertures: Heuristics

1. Initialize: $\hat{X}_0 = X$
2. Main step: at iteration k find $\alpha_k > 0$ and a binary matrix S_k such that

$$\hat{X}_k = \hat{X}_{k-1} - \alpha_k S_k \geq 0$$
$$SPG\left(\hat{X}_k\right) = SPG\left(\hat{X}_{k-1}\right) - \alpha_k$$

3. Termination condition: if $\hat{X}_k = 0$ stop; else, go to Step 2

- To minimize number of apertures we choose maximum possible α_k and corresponding S_k (see [Engel, 2005])

Minimizing Number of Apertures: Heuristics

1. Initialize: $\hat{X}_0 = X$
2. Main step: at iteration k find $\alpha_k > 0$ and a binary matrix S_k such that

$$\hat{X}_k = \hat{X}_{k-1} - \alpha_k S_k \geq 0$$
$$SPG\left(\hat{X}_k\right) = SPG\left(\hat{X}_{k-1}\right) - \alpha_k$$

3. Termination condition: if $\hat{X}_k = 0$ stop; else, go to Step 2

- To minimize number of apertures we choose maximum possible α_k and corresponding S_k (see [Engel, 2005])

Shortcoming of the Sequential Method

- There is often dose discrepancy between FMO and LS solutions
  - some LS methods require rounding fluence maps
  - limited number of apertures are used
- Knowledge of *shape* and *intensity* of apertures are required to model several aspects of IMRT treatment plan
- DAO frameworks have been developed to address these issues

Shortcoming of the Sequential Method

- There is often dose discrepancy between FMO and LS solutions
  - some LS methods require rounding fluence maps
  - limited number of apertures are used
- Knowledge of *shape* and *intensity* of apertures are required to model several aspects of IMRT treatment plan
- DAO frameworks have been developed to address these issues

Direct Aperture Optimization

- Direct aperture optimization (DAO) aims at *directly* finding optimal collection of apertures and their intensities
  - FMO and LS are integrated into a single problem
- In contrast with 3D-conformal radiotherapy where apertures conform to tumor shape in beam's eye-view, in DAO any deliverable aperture by MLC may be used
- We discuss DAO solution methods proposed in [Romeijn et al., 2005] and [Hårdemark et al., 2003]

Direct Aperture Optimization

- Direct aperture optimization (DAO) aims at *directly* finding optimal collection of apertures and their intensities
  - FMO and LS are integrated into a single problem
- In contrast with 3D-conformal radiotherapy where apertures conform to tumor shape in beam's eye-view, in DAO any deliverable aperture by MLC may be used
- We discuss DAO solution methods proposed in [Romeijn et al., 2005] and [Hårdemark et al., 2003]

Direct Aperture Optimization

- Direct aperture optimization (DAO) aims at *directly* finding optimal collection of apertures and their intensities
  - FMO and LS are integrated into a single problem
- In contrast with 3D-conformal radiotherapy where apertures conform to tumor shape in beam's eye-view, in DAO any deliverable aperture by MLC may be used
- We discuss DAO solution methods proposed in [Romeijn et al., 2005] and [Hårdemark et al., 2003]

Column Generation Method: Formulation

- Mathematical formulation for the DAO problem

$$\min G(\mathbf{d})$$

subject to

$$d_v = \sum_{k \in K} \mathcal{D}_{kv} y_k \qquad v \in V$$

$$y_k \geq 0 \qquad k \in K$$

- Notation
  - $\mathbf{y} = (y_k : k \in K)^\top$: vector of aperture intensities
  - $\mathcal{D} = [\mathcal{D}_{kv}]$: matrix of aperture dose deposition coefficients
  - $\mathbf{d} = (d_v : v \in V)^\top$: vector of dose distribution

DAO Solution Challenges

- Naive application of convex optimization techniques to this problem is computationally prohibitive
- K contains a large number of apertures
  - $\mathcal{O}\left(10^{45}\right)$ deliverable apertures per beam angle
- We are interested in sparse solutions
  - clinically reasonable number of apertures (≤ 50 per beam)

DAO Solution Method: Search for Local Minimum

- Necessary optimality conditions for unconstrained problems (i.e., $\nabla F(\bar{\mathbf{y}}) = \mathbf{0}$) can be extended to constrained ones
- If $\bar{\mathbf{y}}$ is a local minimum, then, under some *regularity* conditions, it satisfies *Karush-Kuhn-Tucker (KKT)* conditions
- For a convex problem with affine equality constraints, KKT conditions are necessary and sufficient for global optimality

KKT Optimality Conditions

- If $\bar{\mathbf{y}}$ is a local minimum, then there exist vectors of *Lagrange multipliers* $\bar{\mathbf{u}}, \bar{\mathbf{v}}$ such that this system of equations are satisfied

$$\min F(\mathbf{y})$$

s.t.

$$\begin{aligned} P_\ell(\mathbf{y}) &\leq 0 \qquad \ell \in L_1 \\ Q_\ell(\mathbf{y}) &= 0 \qquad \ell \in L_2 \end{aligned}$$

$$\nabla F(\bar{\mathbf{y}}) + \sum_{\ell \in L_1} \bar{u}_\ell \nabla P_\ell(\bar{\mathbf{y}}) + \sum_{\ell \in L_2} \bar{v}_\ell \nabla Q_\ell(\bar{\mathbf{y}}) = \mathbf{0}$$

$$\begin{aligned} \bar{u}_\ell P_\ell(\bar{\mathbf{y}}) &= 0 & \ell \in L_1 \\ P_\ell(\bar{\mathbf{y}}) &\leq 0 & \ell \in L_1 \\ Q_\ell(\bar{\mathbf{y}}) &= 0 & \ell \in L_2 \\ \bar{u}_\ell &\geq 0 & \ell \in L_1 \end{aligned}$$

KKT Conditions for DAO

- KKT conditions for the DAO problem are as follows:

$$\min G(\mathbf{d})$$

subject to

$$d_v = \sum_{k \in K} \mathcal{D}_{kv} y_k \qquad v \in V$$

$$y_k \geq 0 \qquad k \in K$$

$$\left.\frac{\partial G}{\partial d_v}\right|_{\mathbf{d}=\bar{\mathbf{d}}} - \pi_v = 0 \qquad v \in V$$

$$\sum_{v \in V} \mathcal{D}_{kv} \pi_v - \rho_k = 0 \qquad k \in K$$

$$y_k \rho_k = 0 \qquad k \in K$$

$$y_k \geq 0 \qquad k \in K$$

$$d_v = \sum_{k \in K} \mathcal{D}_{kv} y_k \qquad v \in V$$

$$\rho_k \geq 0 \qquad k \in K$$

- π, ρ: vectors of voxels and apertures Lagrange multipliers

DAO Solution Approach

- We aim at finding $(\bar{\mathbf{y}}, \bar{\mathbf{d}}, \bar{\pi}, \bar{\rho})$ that satisfy KKT conditions
- Due to large number of apertures we cannot incorporate all of them
- We start by considering only a subset of apertures $\hat{K} \subset K$ and sequentially add remaining ones until KKT conditions are all met

DAO Solution Approach: Restricted Problem

- Consider *restricted* DAO problem in which $\hat{K} \subset K$

$$\min G(\mathbf{d})$$

subject to

$$d_v = \sum_{k \in \hat{K}} \mathcal{D}_{kv} y_k \qquad v \in V$$

$$y_k \geq 0 \qquad k \in \hat{K}$$

- This can be solved using a constrained optimization method to obtain $(\mathbf{y}^*, \mathbf{d}^*)$
  - barrier method or projected gradient method

Restricted DAO Problem: KKT Conditions

- Solution $(\mathbf{y}^*, \mathbf{d}^*)$ satisfies KKT conditions for the restricted DAO problem

$$
\begin{aligned}
d_v^* &= \sum_{k \in \hat{K}} \mathcal{D}_{kv} y_k^* && v \in V \\
\pi_v^* &= \left. \frac{\partial G}{\partial d_v} \right|_{\mathbf{d}=\mathbf{d}^*} && v \in V \\
\rho_k^* &= \sum_{v \in V} \mathcal{D}_{kv} \pi_v^* && k \in \hat{K} \\
y_k^* \rho_k^* &= 0 && k \in \hat{K} \\
y_k^* \geq 0 \,&, \rho_k^* \geq 0 && k \in \hat{K}
\end{aligned}
$$

- We then construct solution $\bar{\mathbf{y}}$ as $\bar{y}_k = \begin{cases} y_k^* & k \in \hat{K} \\ 0 & k \in K \setminus \hat{K} \end{cases}$

DAO Solution Approach: KKT Conditions

- We substitute $\bar{\mathbf{y}}$ in KKT conditions of original DAO problem

$$
\begin{aligned}
\bar{d}_v &= \sum_{k \in K} \mathcal{D}_{kv} \bar{y}_k && v \in V \\
\bar{\pi}_v &= \left. \frac{\partial G}{\partial d_v} \right|_{\mathbf{d} = \bar{\mathbf{d}}} && v \in V \\
\bar{\rho}_k &= \sum_{v \in V} \mathcal{D}_{kv} \bar{\pi}_v && k \in K \\
\bar{y}_k \bar{\rho}_k &= 0 && k \in K \qquad \textit{why?} \\
\bar{y}_k &\geq 0 && k \in \hat{K} \\
\bar{y}_k &= 0 && k \in K \setminus \hat{K} \\
\bar{\rho}_k &\geq 0 && k \in \hat{K} \qquad \textit{why?} \\
\bar{\rho}_k &\geq 0 && k \in K \setminus \hat{K} \qquad ?
\end{aligned}
$$

DAO Solution Approach: Pricing Problem

- To ensure if $\bar{\rho}_k \geq 0$ for $k \in K$ we formulate and solve the *pricing problem*

$$\min_{k \in K} \bar{\rho}_k = \sum_{v \in V} \mathcal{D}_{kv} \bar{\pi}_v$$

- Aperture $k \in K$ consists of a collection, A_k, of exposed beamlets $i \in I$

$$\mathcal{D}_{kv} \approx \sum_{i \in A_k} D_{iv}$$

- Pricing problem is reformulated using beamlet dose deposition coefficients

$$\min_{k \in K} \sum_{i \in A_k} \sum_{v \in V} D_{iv} \bar{\pi}_v$$

DAO Solution Approach: Pricing Problem

- To ensure if $\bar{\rho}_k \geq 0$ for $k \in K$ we formulate and solve the *pricing problem*

$$\min_{k \in K} \bar{\rho}_k = \sum_{v \in V} \mathcal{D}_{kv} \bar{\pi}_v$$

- Aperture $k \in K$ consists of a collection, A_k, of exposed beamlets $i \in I$

$$\mathcal{D}_{kv} \approx \sum_{i \in A_k} D_{iv}$$

- Pricing problem is reformulated using beamlet dose deposition coefficients

$$\min_{k \in K} \sum_{i \in A_k} \sum_{v \in V} D_{iv} \bar{\pi}_v$$

DAO Solution Approach: Pricing Problem

- To ensure if $\bar{\rho}_k \geq 0$ for $k \in K$ we formulate and solve the *pricing problem*

$$\min_{k \in K} \bar{\rho}_k = \sum_{v \in V} \mathcal{D}_{kv} \bar{\pi}_v$$

- Aperture $k \in K$ consists of a collection, A_k, of exposed beamlets $i \in I$

$$\mathcal{D}_{kv} \approx \sum_{i \in A_k} D_{iv}$$

- Pricing problem is reformulated using beamlet dose deposition coefficients

$$\min_{k \in K} \sum_{i \in A_k} \sum_{v \in V} D_{iv} \bar{\pi}_v$$

DAO Solution Approach: Pricing Problem

- It finds aperture k with most negative Lagrange multiplier at $\bar{\mathbf{y}}$
  - $\bar{\rho}_k$: rate of change in G as intensity of aperture k increases (reduced gradient)

$$\min_{k \in K} \; \bar{\rho}_k = \sum_{i \in A_k} \underbrace{\sum_{v \in V} D_{iv} \bar{\pi}_v}_{\text{beamlet i's}}$$

  - given reduced gradient of all beamlets, it finds collection of beamlets that
    - forms a deliverable aperture
    - has most negative cumulative reduced gradient

5	-1	-2	-3	1
-1	3	-1	-3	2
-1	2	-5	-1	1
1	1	2	1	3

Pricing Problem: Pricing Problem

- Pricing problem can be solved for individual beam angles $b \in B$

$$\min_{k \in K_b} \bar{\rho}_k$$

- Pricing problem can be separated over beamlet rows
  - finding a *sub sequence* of beamlets with most negative cumulative reduced gradient

2	-1	3	-2	1	-2	2	-1

Pricing Problem: Minimum Subsequent Sum

- It can be solved by a single pass over bealmests in a row $i = 1, \ldots, N$

1. Initialize: $b_i = \sum_{v \in V} D_{iv} \pi_v$ for $i \in I$, $minSoFar = 0$, $minEndingHere = 0$
2. Main step: For $i = 1 : N$

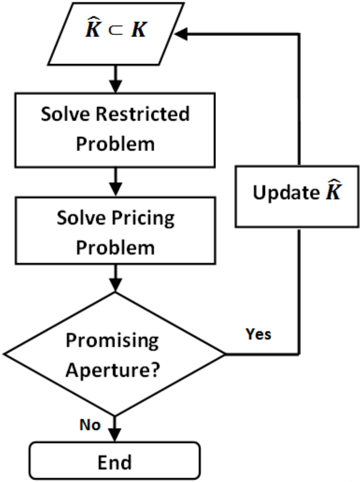
$$minEndingHere = \min \left\{ 0, minEndingHere + b_i \right\}$$
$$minSoFar = \min \left\{ minEndingHere, minSoFar \right\}$$

3. Output: $minSoFar$

- The pricing problem can be extended to incorporate additional MLC hardware constraints

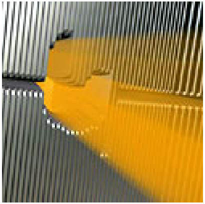
Column Generation Method

- *Column generation method* for DAO solves restricted and pricing problems iteratively
  - common approach to solve large-scale optimization problems
- It can be terminated if no promising aperture exists or if current solution is satisfactory



Leaf Refinement Problem

- Given a fixed number of apertures, the *leaf refinement problem* aims at finding their optimal leaf positions as well as their intensities
  - see [Hårdemark et al., 2003, Cassioli and Unkelbach, 2013]
- Major difference from column generation approach is that the number of apertures is fixed



Leaf Refinement Method: Dose Deposition

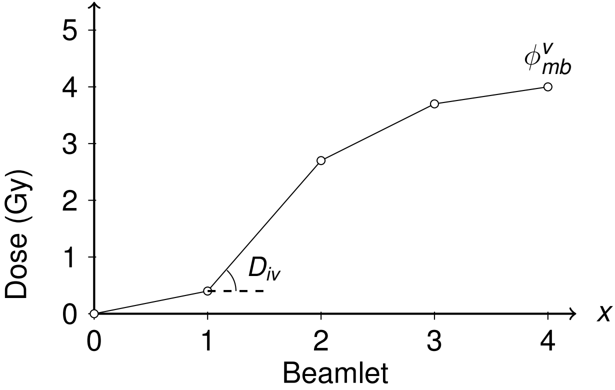
- Expressing dose deposited in voxel $v \in V$ in terms of aperture intensities and leaf positions

$$d_v\left(\mathbf{x}^{(\mathbf{l})}, \mathbf{x}^{(\mathbf{r})}, \mathbf{y}\right) \equiv \sum_{k \in K} y_k \sum_{m \in M} \left(\phi^v_{mb(k)}\left(x^{(r)}_{mk}\right) - \phi^v_{mb(k)}\left(x^{(l)}_{mk}\right)\right)$$

- Notation
  - $\mathbf{y} = (y_k : k \in K)^\top$: vector of aperture intensities
  - $\mathbf{x}^{(\mathbf{l})} = \left(x^{(l)}_{mk} : m \in M, k \in K\right)^\top$: vector of left leaf positions
  - $\mathbf{x}^{(\mathbf{r})} = \left(x^{(r)}_{mk} : m \in M, k \in K\right)^\top$: vector of right leaf positions
  - $\phi^v_{mb}(x)$: dose deposited in voxel $v \in V$ under unit intensity from row $m \in M$ in beam angle $b \in B$ when interval $[0, x]$ is exposed

Leaf Refinement Method: Dose Deposition

- Given beamlet dose deposition coefficient matrix $[D_{iv}]$, ϕ^v_{mb} can be approximated using a piecewise-linear function



Leaf Refinement Problem: Formulation

- Mathematical formulation

$$\min G(\mathbf{d}(\mathbf{z}))$$

subject to

$$\begin{aligned} \mathbf{A}\mathbf{x} &\leq \mathbf{0} \\ H(\mathbf{d}(\mathbf{x},\mathbf{y})) &\leq \mathbf{0} \;\; \rightarrow F_\ell(\mathbf{d}(\mathbf{z})) \leq 0 \qquad \ell \in L \\ \mathbf{y} &\geq \mathbf{0} \end{aligned}$$

- Notation
  - $\mathbf{d} = (d_v : v \in V)^\top$: vector of dose distribution
  - $\mathbf{y} = (y_k : k \in K)^\top$: vector of aperture intensities
  - $\mathbf{z} = \left(\mathbf{x}^{(\mathbf{l})}\ \mathbf{x}^{(\mathbf{r})}\ \mathbf{y}\right)^\top$: vector of all variables
  - $\mathbf{A}$: constraint matrix of leaf positions

Leaf Refinement Problem: Solution Approach

- *Sequential quadratic programming* (SQP) can be used
- SQP aims at finding a solution that satisfies KKT conditions

$$\begin{aligned} \nabla G(\mathbf{z}) + \sum_{\ell \in L} u_\ell \nabla F_\ell(\mathbf{z}) &= \mathbf{0} \\ u_\ell F_\ell(\mathbf{z}) &= 0 \qquad\qquad \ell \in L \\ \mathbf{u} &\geq \mathbf{0} \end{aligned}$$

- One can use Newton method to solve this system
  - Lagrangian is defined as $\mathcal{L} \equiv G + \sum_{\ell \in L} u_\ell F_\ell$
  - we let

$$\nabla F^\top = \begin{pmatrix} \nabla F_1^\top \\ \vdots \\ \nabla F_L^\top \end{pmatrix}$$

Sequential Quadratic Programming

- Newton method at iteration k requires to solve

$$\begin{pmatrix} \nabla^2 \mathcal{L}_{(k)} & \nabla F_{(k)}^\top & & \\ u_1 \nabla F_{1(k)}^\top & F_{1(k)} & 0 & \cdots & 0 \\ u_2 \nabla F_{2(k)}^\top & 0 & F_{2(k)} & \cdots & 0 \\ u_L \nabla F_{L(k)}^\top & 0 & 0 & \cdots & F_{L(k)} \end{pmatrix} \begin{pmatrix} \mathbf{z} - \mathbf{z}_{(k)} \\ \mathbf{u} - \mathbf{u}_{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla G_{(k)} + \sum_{\ell \in L} u_{\ell(k)} \nabla F_{(k)} \\ u_{1(k)} F_{1(k)} \\ \vdots \\ u_{L(k)} F_{L(k)} \end{pmatrix}$$

this reduces to

$$\nabla^2 \mathcal{L}_{(k)} \left(\mathbf{z} - \mathbf{z}_{(k)} \right) + \nabla F_{(k)}^\top \mathbf{u} = -\nabla G_{(k)}$$

$$u_\ell \left(\nabla F_{\ell(k)}^\top \left(\mathbf{z} - \mathbf{z}_{(k)} \right) + F_{\ell(k)} \right) = 0 \qquad \ell \in L$$

Sequential Quadratic Programming

- Along with $\mathbf{u} \geq \mathbf{0}$ these are also KKT conditions for the following *quadratic programming* (QP) problem

$$\min G_{(k)} + \nabla G_{(k)}^\top \mathbf{v} + \frac{1}{2}\mathbf{v}^\top \nabla^2 \mathcal{L}_{(k)} \mathbf{v}$$

subject to

$$\nabla F_{\ell(k)}^\top \mathbf{v} + F_{\ell(k)} \leq 0 \qquad \ell \in L$$

in which we substituted $\mathbf{v} = \mathbf{z} - \mathbf{z}_{(k)}$

Sequential Quadratic Programming

- One can alternatively solve this QP problem at iteration k of Newton method
  - objective function is quadratic approximation of G plus curvature of constraints at $\mathbf{z} = \mathbf{z}_k$
  - constraints are linear approximation of F_ℓ ($\ell \in L$) at $\mathbf{z} = \mathbf{z}_k$
- The QP problem requires computing $\nabla^2 \mathcal{L}_{(k)}$
  - computationally expensive
  - may not be positive definite

Sequential Quadratic Programming: BFGS Update

- To overcome this issue *quasi-newton* method is employed

$$\nabla^2 \mathcal{L}_{(k)} \approx B_{(k)} \succeq 0$$

- Positive definite approximations of Hessian using (Broyden-Fletcher-Goldfarb-Shanno) BFGS update

$$\mathbf{B}_{(k+1)} = \mathbf{B}_{(k)} + \frac{\mathbf{q}_{(k)}\mathbf{q}_{(k)}^\top}{\mathbf{q}_{(k)}^\top \mathbf{p}_{(k)}} - \frac{\mathbf{B}_{(k)}\mathbf{p}_{(k)}\mathbf{p}_{(k)}^\top \mathbf{B}_{(k)}}{\mathbf{p}_{(k)}^\top \mathbf{B}_{(k)}\mathbf{p}_{(k)}}$$

$$\mathbf{p}_k = \mathbf{z}_{(k+1)} - \mathbf{z}_{(k)} \qquad \mathbf{q}_k = \nabla \mathcal{L}_{(k+1)} - \nabla \mathcal{L}_{(k)}$$

Leaf Refinement Method: SQP

1. Initialization: select initial variables $\mathbf{z}_{(0)}$, lagrange multipliers $\mathbf{u}_{(0)}$, and Hessian p.d. approximation $\mathbf{B}_{(0)}$
2. Main Step: at iteration k solve the QP problem

$$\min_{\mathbf{v}} \; G_{(k)} + \nabla G_{(k)}^{\top}\mathbf{v} + \frac{1}{2}\mathbf{v}^{\top}\mathbf{B}_{(k)}\mathbf{v}$$

subject to

$$F_{\ell(k)} + \nabla F_{\ell(k)}^{\top}\mathbf{v} \leq 0 \qquad \ell \in L$$

3. Termination condition: if $\|\mathbf{v}^*\| < \epsilon$, stop; otherwise, $\mathbf{z}_{(k+1)} = \mathbf{z}_{(k)} + \mathbf{v}^*$, update $\mathbf{B}_{(k+1)}$ using BFGS method and go to Step 2

Summary: DAO

- DAO aims at directly solving for aperture shape and intensities
- DAO plans employ fewer apertures and shorter beam-on times compared to two-stage method to obtain similar dose conformity
  - see [Ludlum and Xia, 2008, Men et al., 2007]
- We discussed two major DAO approaches
  - column generation method
  - leaf refinement method

References I

- Ahuja, R. K. and Hamacher, H. W. (2005).
  A network flow algorithm to minimize beam-on time for unconstrained multileaf collimator problems in cancer radiation therapy.
  Networks, 45(1):36–41.

- Aleman, D. M., Glaser, D., Romeijn, H. E., and Dempsey, J. F. (2010).
  Interior point algorithms: guaranteed optimality for fluence map optimization in IMRT.
  Physics in Medicine and Biology, 55(18):5467–5482.

- Baatar, D., Hamacher, H. W., Ehrgott, M., and Woeginger, G. J. (2005).
  Decomposition of integer matrices and multileaf collimator sequencing.
  Discrete Applied Mathematics, 152(1-3):6–34.

- Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2006).
  Nonlinear programming: theory and algorithms.
  Wiley-interscience.

References II

- Cassioli, A. and Unkelbach, J. (2013).
  Aperture shape optimization for IMRT treatment planning.
  Physics in Medicine and Biology, 58(2):301–318.

- Engel, K. (2005).
  A new algorithm for optimal multileaf collimator field segmentation.
  Discrete Applied Mathematics, 152(1):35–51.

- Hårdemark, B., Liander, A., Rehbinder, H., and Löf, J. (2003).
  Direct machine parameter optimization with raymachine® in pinnacle3®.
  RaySearch White Paper. Stockholm, Sweden: RaySearch Laboratories AB.

References III

- Ludlum, E. and Xia, P. (2008).
  Comparison of IMRT planning with two-step and one-step optimization: a way to simplify IMRT.
  Physics in Medicine and Biology, 53(3):807–821.
- Men, C., Romeijn, H. E., Taşkın, Z. C., and Dempsey, J. F. (2007).
  An exact approach to direct aperture optimization in IMRT treatment planning.
  Physics in Medicine and Biology, 52(24):7333–7352.
- Romeijn, H., Ahuja, R., Dempsey, J., and Kumar, A. (2005).
  A column generation approach to radiation therapy treatment planning using aperture modulation.
  SIAM Journal on Optimization, 15(3):838–862.

References IV

Webb, S. (2001).
Intensity-modulated radiation therapy.
Institute of Physics Publishing.